連続値特徴量の効率的なパターンマイニングに関する研究

修士学位論文要約（令和３年３月）

連続値特徴量の効率的なパターンマイニングに関する研究

對馬 裕樹
指導教員：篠原 歩

Efficient pattern mining on numerical features

Yuki TSUSHIMA
Supervisor: Ayumi SHINOHARA

Linear models are simple and interpretable, and we can improve their accuracy by using patterns, i.e., combinations of features. For further improvement of their accuracy, it is desirable to keep numerical values as they are rather than binarizing them. However, there are no efficient mining algorithms for numerical features focused on classification problems. In this paper, we propose an algorithm for constructing linear models by mining patterns without binarization. Our algorithm mines significantly faster than other algorithms and can achieve high accuracy.

1. はじめに

機械学習分野において**解釈性**の高い分類モデルの構築は重要な課題である．例えば，線形モデルは解釈性の高いモデルであり，さらに特徴量の組み合わせ，すなわち**パターン**を特徴量として用いることで分類精度の向上が期待できる．データセット中のクラスを上手く識別するパターンの抽出は，モデルの精度向上だけではなく解釈性の向上にも寄与する．抽出されるパターンに関する条件となる**制約**を設けた**制約パターン** [3) 1)] は，あるクラスの特徴を捉え，他のクラスのデータとの差異を識別する．データセット中のパターンの出現率である**サポート**が制約としてよく用いられるが，連続値データセットに対するサポートを考えるためには連続値特徴量の二値化を行う必要がある．しかしながら，二値化を行うことでデータセット中のクラスを識別するための重要な情報が欠落したり，特徴量数が増加することでマイニングにかかる時間が増大してしまう場合がある．

Tatti [6)] は連続値データセットに対して *copula support* を定義した．Copula support を用いることで，二値化を行わずに，連続値データセットでの制約パターンマイニング問題を定義することができる．

2. 本研究の成果

制約パターンの探索空間は特徴量次元数に関して指数的に大きく，copula support を用いた制約パターンマイニング問題を高速に解くためには効率的な枝刈り探索を行うことが必要である．本研究では，copula support を用いた制約パターンマイニングにおいて，探索木を用いた探索に効果的な枝刈りと探索順序の動的化を導入することで高速に動作するマイニングアルゴリズムを提案する．提案手法では，データセット中の負ラベルデータのサポートの下界を用いて，未探索の任意のアイテムセットを追加したパターンのサポートの上界を求めることで，いかなるアイテムを加えたパターンも制約を満たさないようなケースを早期に発見し，探索空間の枝刈りを行う．また，枝刈りを早期に適応するために，各探索ステップでサポートに基づいてアイテムの探索順序を動的に決定する．さらに，アイテム探索順序の動的化を効率的に行うために，双方向連結リストを用いてデータセットを管理する．上記に加え，特徴量の小ささを示す**ネガティブアイテム**を新たな特徴量として導入する．ネガティブアイテムは，分類精度だけでなく解釈性の向上にも寄与する．

3. 実験

LIBSVM データセット [2)]，および UCI repository [4)] で公開されているデータセットを用いて，提案手法によって得られるパターンを用いて生成した分類モデルの精度と学習時間を既存手法と比較する．既存手法として，ロジスティック回帰 (LR)，線形カーネルサポートベクターマシン (SVM)，safe pattern pruning(SPP) [5)] を用いる．SPP は，バイナリデータセットを用いた L1 正則化付き最適化問題において，最適解に含まれないパターンのみを事前に取り除くことで予測モデルを構築する際に有用なパターンのみを発見する手法である．SPP は二値データセットを入力値とする手法であるため，事前に連続値データセットをビニングによって二値データセッ

表 1: 提案手法と既存手法を用いた分類モデルの精度比較

Dataset	提案手法	SPP	LR	SVM
wine	0.964	**0.974**	0.938	0.929
breast_cancer	0.957	**0.959**	0.950	0.954
fourclass	0.778	**0.942**	0.754	0.751
haberman	0.620	**0.701**	0.642	0.631
liver-disorders	0.696	**0.756**	0.742	0.735
ctg_class	**0.823**	0.807	0.772	0.765
ctg_nsp	0.905	**0.914**	0.893	0.892
faults	**0.712**	0.706	0.686	0.686
magic	**0.840**	0.831	0.781	0.779
segmentation	**0.969**	0.959	0.923	0.917
transfusion	0.733	**0.744**	0.700	0.688

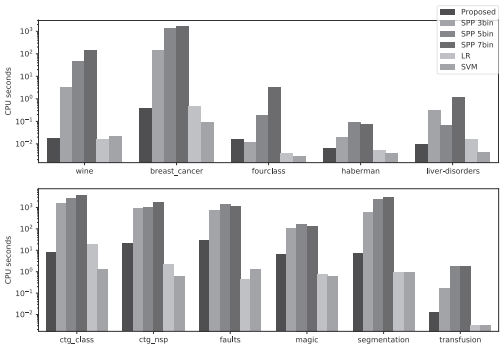

図 1: 提案手法と既存手法を用いた分類モデルの学習時間の比較

トに変換する前処理を行う．

表 1 に，各データセットについて各手法を用いたときの F 値を示す．SPP と比較して提案手法は，ctg_class データセットや，faults データセット，magic データセットといった，特徴量数が多いデータセットにおいて高いパフォーマンスが得られた．

図 1 に各手法を用いたときの計算時間について示す．図より，全てのデータセットについて，提案手法が SPP より高速であることがわかる．特に，特徴量数の多いデータセットにおいて，提案手法は SPP と比較して最大で 1000 倍ほど高速である．これは，SPP においては，ビニングによってパターンの候補となる特徴量数が大きく増加してしまうためである．提案手法は，SPP に匹敵する精度を示す分類モデルを，より高速に構築することができると結論付けられる．

breast_cancer データセットに対して提案手法を用いることで得られたパターンについて，*malignant* クラスの F 値に最も寄与した 3 つのパターンは以下であった．

$$\begin{aligned} &\textit{high mean texture} \wedge \textit{high compactness error} \\ &\qquad \wedge \textit{high worst fractal dimension}, \\ &\textit{high mean concavity} \wedge \textit{high concave points error} \\ &\qquad \wedge \textit{low symmetry error}, \\ &\textit{high mean texture} \wedge \textit{high mean compactness}. \end{aligned}$$

これらのうち，“low” で始まるものはネガティブアイテムである．2 つ目のパターンを見ると，*mean concavity* と *concave points error* が共に大きく，*symmetry error* が小さいデータポイントは *malignant* クラスに分類されやすいということがわかる．ネガティブアイテムを用いることで，分類精度に寄与する連続値特徴量間の相関関係を容易に捉えることができる．

文献

1) R. J. Bayardo, R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense databases. *Data mining and knowledge discovery*, Vol. 4, No. 2, pp. 217–240, 2000.
2) C. C. Chang and C. J. Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, Vol. 2, No. 3, pp. 1–27, 2011.
3) G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proc. KDD 1999*, pp. 43–52. ACM, 1999.
4) D. Dua and C. Graff. UCI machine learning repository, 2017.
5) K. Nakagawa, S. Suzumura, M. Karasuyama, K. Tsuda, and I. Takeuchi. Safe pattern pruning: An efficient approach for predictive pattern mining. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1785–1794, 2016.
6) N. Tatti. Itemsets for real-valued datasets. In *IEEE International Conference on Data Mining, ICDM 2013*, pp. 717–726. IEEE, 2013.